Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Karolína Kuchyňová
Název práce	Ověřování identity uživatele založené na behavoirálních charakteristikách
Rok odevzdání	2020
Studijní program	Informatika **Studijní obor** Umělá inteligence

Autor posudku	Ladislav Peška **Role** Oponent
Pracoviště	KSI

Text posudku:

Práce se zabývá problematikou identifikace, případně autentizace uživatele na základě jeho behaviorálních charakteristik bez použití speciálního hardwaru (tj. sledování akcí myši a klávesnice).

Práce postupně rozebírá existující přístupy v dané oblasti, definuje akce, které mohou být použity pro autentizaci, popisuje dataset, který byl v rámci práce sestaven, testuje použitelnost jednotlivých příznaků a nakonec aplikuje několik standardních machine learning (ML) algoritmů pro autentizaci uživatelů.

Textová část práce je poměrně rozsáhlá - v některých částech možná až zdlouhavá, ale na druhou stranu tato zdlouhavost umožňuje číst pouze vybrané části práce (například pouze vyhodnocení některých příznaků), takže ji celkově hodnotím spíše jako přínos. Hlavní přínos práce vidím v sestavení rozsáhlého datasetu, který v dané oblasti doposud citelně chyběl. Autorka dobře argumentuje zvolenou platformu i průběh testování - velkou výhodou je doba sběru dat, která umožňuje zkoumání časových závislostí v chování uživatelů. Na druhou stranu mohlo být více péče věnováno samotnému výběru zaznamenávaných akcí. Například mi chybí odůvodnění pro zvolenou periodu snímkování pohybů myši, nebo informace o průměrném čase, po který byly pohyby myši zaznamenávané. (Jak často dochází k vyvolání mouseMove události při souvislém pohybu? Existují zde rozdíly mezi prohlížeči?) Obdobně pak mohlo být důležité zaznaménávat parametry webové stránky, které mohly chování uživatele ovlivnit (například celkové rozměry stránky, rozlišení obrazovky, přítomnost různých GUI elementů apod.).

Práce obsahuje poměrně rozsáhlou analýzu datasetu i jednotlivých testovaných příznaků. Některé dílčí otázky sice zůstávají nezodpovězeny (například jaký je timespan histogram pro jednotlivé uživatele), nicméně základní přehled o shromážděných datech práce obsahuje. Za poněkud nevhodně zvolenou považuju metriku \delta_{k,n}, která může obsahovat dělení nulou (autorka sama uvádí jeden takový příklad na str. 81). Jedná se o metriku používanou v obdobných studiích? Také ne zcela souhlasím se závěry v kapitole 4.2.1 a 4.2.2 - i přesto, že je v jednotlivých session zaznamenáno jen minimum úhozů kláves, je možné uvažovat například odchylky od průměrných hodnot pro jednotlivé klávesy nebo skupiny kláves např. v jednoduchém bias modelu.

Jako slabší hodnotím použití ML algoritmů v kapitole 5. Především absencí hyperparameter tuningu mohlo dojít k poměrně zásadnímu ovlivnění výsledků. Použité algoritmy jsou standardní, odpovídající přehledu v related work bez dodatečné invence na straně autorky.

Osobně by mě zajímala například možnost využití rekurentních neuronových sítí pro tento druh dat. I přes popsané nedostatky se však jedná o velmi zajímavou práci, kterou mohu doporučit k obhajobě.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 31. August 2020 **Podpis**